ATD4MA:多属性数据的联合真值发现方法 *

何　杰，卢　菁†，邵　清，刘　丛
(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘　要：目前已提出的真值发现方法无法解决对象由多个单值属性与多值属性共同组成的情况，若将这些属性拆分后分别处理则会破坏属性间原有的关联，导致计算结果不准确。提出一种多属性数据的联合真值发现方法 ATD4MA，将对象各观察值通过遗传算法中的染色体进行建模，针对问题特性对群体初始化算法和染色体基本动作进行改进，控制染色体的演化行为对各属性进行约束，以各对象的真值染色体与各数据源提供的观察值染色体间的差异加权和达到最小为目标建立优化模型，解决了对象包含多个属性的真值发现问题。在两个真实数据集上的实验，证明了提出方法的正确性和有效性。

关键词：真值发现；数据相关性；单值属性；多值属性；遗传优化算法
中图分类号：TP311　　doi: 10.19734/j.issn.1001-3695.2018.12.0885

ATD4MA: associated truth discovery method for multi-attribute data

He Jie, Lu Jing, Shao Qing, Liu Cong
(*School of Optical-electrical & Computer Engineering*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*)

Abstract: The current truth discovery method cannot solve the case where the object is composed of many single-valued attributes and multi-valued attributes. Separate processing of these attributes will destroy the original association between attributes, resulting in inaccurate results. This paper proposed an associated truth discovery method for multi-attribute data (ATD4MA). It modeled the observation values of the object using the chromosomes in the genetic algorithm. Then it improved the population initialization algorithm and the basic action of the chromosome according to the problem characteristics. By controlling the evolution behavior of chromosomes, it established the optimization model to minimize the weighted sum of difference between the truth-value chromosomes of each object and the observed values provided by each data source. Therefore it solved the problem of truth discovery where the object contains multiple attributes. Experiments on two real data sets show the correctness and effectiveness of the proposed method.

Key words: truth discovery; data dependency; single truth attribute; multiple truth attribute; genetic optimization algorithm

0　引言

随着互联网的快速发展，Web 已经成为人们获取信息的主要手段之一。与此同时，数据冲突问题也日益凸显，不同数据源会对同一对象赋予截然不同的描述，大量错误、过时、不完整、虚假的信息混杂于网络之中，这使得用户难以辨别正确的结果。这些错误的信息不仅没有任何利用价值，且使得整个 Web 数据源提供的信息成为一个冲突集，给用户带来误导甚至造成巨大的损失。例如，各网站对电影《芳华》的演员列表和总时长两个属性提供了不同的描述信息，如表 1 所示。其中，豆瓣提供了正确的演员列表、错误的电影总时长；优酷提供了不完整的演员列表、正确的电影时长；而电影天堂提供的描述中存在错误的信息，误将该电影的编剧记录为演员。

由此可以看出，不同网站对于同一对象的描述存在大量冲突信息，这些冲突信息可能由人为的粗心大意、信息长时间未更新或语义分析不正确等原因造成，这不仅对用户的查询起到误导作用，且使得原本正确的信息无法被辨别，给用户带来巨大的不便。如何在有冲突的数据集中找到值得信赖的信息，学术界称之为真值发现问题。

表 1　各网站提供的电影信息

Table 1　Movie information provided by each website

数据源	演员列表	电影时长
豆瓣	黄轩,苗苗,钟楚曦,杨采钰,李晓峰,王天辰,王可如,隋源,张仁博,苏岩,张国立,赵立新	146 min
优酷	黄轩,苗苗,钟楚曦，李晓峰	136 min
1905 电影	黄轩,苗苗,钟楚曦,杨采钰,李晓峰	130 min
电影天堂	严歌苓,黄轩,苗苗,钟楚曦,杨采钰,李晓峰,王天辰,王可如,隋源	136 min

早期的解决方法采取投票机制，即出现最多次数的信息被认为是最可信的，此方法认为数据源提供了相同可信度的信息，忽略了数据源质量上的差异。目前真值发现处理方法可分为两类：a)利用迭代机制，根据高质量的数据源可能提供高可信的数据，大量高可信数据可能源于高质量数据源的原理，反复迭代更新各数据源权值与对象真值集合直至算法达到收敛状态; b)基于概率，通过建立概率模型来推断观察值为真值的可能性，从而确定真值集合。

相关研究经历了两个阶段，第一阶段为单真值发现。文

收稿日期：2018-12-27；**修回日期：**2019-02-25　**基金项目：**国家自然科学基金青年基金项目（61703278）
作者简介：何杰（1992-），男，甘肃兰州人，硕士研究生，主要研究方向为真值发现、数据处理；卢菁（1976-），女（通信作者），江苏东台人，讲师，博士，主要研究方向为数据一致性、真值发现（jing.lu@usst.edu.cn）；邵清（1970-），女，上海人，副教授，博士，主要研究方向为数据集成、容错计算；刘丛（1983-），男，山东高唐人，讲师，博士，主要研究方向为机器学习.

献[1]首次定义真值发现问题，提出一种基于贝叶斯概率模型的启发式算法 TruthFinder，通过无监督的迭代机制联合计算数据源质量和真值可信度。在此基础之上为减少各负面因素对数据源权值计算的影响，文献[2]基于概率模型，提出事实难易程度的概念，以提高各数据源可信度。文献[3~6]提出了数据间各种复制关系及关联关系的检测及处理方法，以减少数据间的复制关系对数据源权值计算精确度的影响。文献[7]考虑到数据中普遍存在的长尾现象，提出基于置信度的方法，减少长尾现象对结果精确性的影响。文献[8,9]认为同一数据源下不同知识领域的数据应有不同的权值分配，提出基于知识领域的权值分配方法。文献[1～9]可以统称为真值发现问题解决方案的第一个阶段，因为它们只能处理单真值单属性的问题。

第二阶段的研究为多真值发现。文献[10]提出一种基于贝叶斯的半监督学习方法，通过拟定的值与观察值之间的相似性通过迭代算法直到收敛来计算真值结果集。文献[11]通过构建概率图模型 LTM，假设数据源的准确率和查全率服从 Beta 分布，从而得到属性值为真的概率，然后根据计算得到的概率值与之前设定好阈值间大小关系得到最终的真值集。然而，此方法在阈值的设定规则与最终属性值的选择策略上并没有给出严格的定义，不同的阈值设定会对结果的精确性造成不同程度的影响。文献[12]提出数据相似性概念，建立优化模型避免阈值的设定。文献[13]基于严格的数学推导，对无监督迭代型算法最终是否可达到收敛给出了证明。

文献[14]首次提出并解决真值发现问题中对象存在多个属性的问题，但该方法只考虑到多属性单真值的问题，并未涉及多属性多真值的问题。当对象包含一个或多个单值属性的同时也包含一个或多个多值属性时，目前已有的真值发现方法无法解决此种情况。以表 1 中电影数据为例，电影时长属性为单值，演员列表属性为多值。若仅通过现有的真值发现方法来解决此问题，可尝试两种方案，分析如下：

a)将所有属性视为一个整体，通过多真值发现方法模型进行求解。理论上可以求得最相似的真值结果集，但无法避免单值属性最终被赋予多个值的情况。

b)将属性拆分为两个子问题，分别通过单真值发现方法和多真值发现方法进行计算。该方法刻意地将属性进行拆分，破坏了对象属性间的关联，导致在数据源权值计算上出现偏差，从而得到不精确的结果。

综上，目前已提出方法无法同时进行多个属性的联合处理。若按照之前方法则只能将属性拆分为单个属性去套用只能处理一个属性的算法，这完全忽略了属性间相关性的存在，且拆分后很难有一个合理的合并方案，因此需提出一种整体求解的方法，确保不破坏属性相关性。本文提出一种联合真值发现方法，首先将每条记录用遗传算法中染色体的形式表示，并在遗传算法原有流程的基础上针对本问题进行改进，同时可以克服多属性情况下目标函数存在多个局部最优解的问题，在不破坏属性相关性的前提下计算出对象真值结果集。该方法不仅可用于一般的单真值与多真值发现，且在该研究领域取得一大突破，即可解决对象同时包含多个单值与多值属性的真值发现问题。本文主要贡献总结如下：

a)本文提出联合真值发现问题，证明了多属性数据中各属性间存在相关性的事实，以此建立出合理的真值求解模型，确保在计算过程中不会破坏对象属性间相关性。

b)提出数据差异性概念，定义损失函数用于评价两数据间的冲突程度。以数据整体差异性最小为目标建立优化模型，无须设置阈值及制定选择策略，避免了人为设定造成的影响。

c)一定程度上消除了长尾现象对数据源权值计算的影响。将数据源权值除以该数据源提供对象的数量，平滑了各数据源提供对象数量上的差异，以进一步提升计算结果的精准度。

d) 提出 ATD4MA(associated truth discovery for multi-attribute data)算法，利用遗传算法中染色体对多属性多真值发现问题进行联合建模，并改进了算法流程。通过改变群体初始化算法与染色体演化时的基本动作对其变化后的内部基因特征进行控制，从而约束属性值的数量，避免单值属性出现多个解的错误情况，并通过优胜劣汰的原则对染色体种群进行演化从而逐步逼近全局最优解，从目标函数的多个局部最优解中找到全局最优解，在线性时间内计算出各数据源最优权值分配与最终多真值结果集，解决了多属性对象的真值发现问题。

e)通过在两个真实数据集上的实验，验证了本文提出方法的准确性和有效性。

1 系统框架

1.1 相关定义

定义 1 对象可能值集。所有数据源为该对象提供属性值的集合。为不破坏对象属性值原本的顺序，在多值属性值的放置顺序上本文采取投票统计的方式进行排序，即将多值属性的某个位置上出现的各值进行统计，出现次数最多值放置于当前位置。在单值属性的处理上，不涉及顺序问题，为方便之后对照将其中名称按照字母序排列，数值按照大小排列。

例 1 表 1 所示的电影数据中，以演员属性为例，经过统计后发现各对象描述数据中第一位出现“黄轩”的次数最大，其对象可能值集中第一位置应放“黄轩”。故《芳华》对应的可能值集为{黄轩,苗苗,钟楚曦,杨采钰,李晓峰,王天辰,王可如,隋源,张仁博,苏岩,张国立,赵立新,严歌苓, 130 min, 136 min, 146 min}。

定义 2 对象观察值染色体。表示了数据源提供观察值在对象可能值集上的分布情况，可直接视为遗传优化算法中的标准染色体，其长度为该对象可能值集的长度，每个基因初始值为 0，取值范围为{0,1}。当观察值提供了可能值集上的第 i 个值时，则将该染色体中第 i 个基因的值标记为 1。

例 2 若得到一组电影数据的观察值为{黄轩,苗苗,钟楚曦,杨采钰,李晓峰,王天辰,王可如,隋源,136 min}，则其对象观察值染色体为{1,1,1,0,1,1,1,1,0,1,0,0,0,0,1,0}。

定义 3 观察值属性染色体。对象观察值染色体的子染色体，对应该对象中的某个属性。

例 3 例 2 中描述电影时长属性的观察值染色体为{0,1,0}。

定义 4 对象真值染色体。用来表示某对象的真值集合，其长度为该对象可能值集的长度，每个基因初始值全为 0，取值范围为{0,1}。当真值集合中存在该对象可能值集上的第 i 个值时，则将该染色体中第 i 个基因的值标记为 1。

例 4 若电影《芳华》的真值集合为{黄轩,苗苗,钟楚曦,杨采钰,李晓峰,王天辰,王可如,146 min}，则该电影对象的真值染色体为{1,1,1,0,1,1,1,1,0,0,0,0,0,0,0,1}。

定义 5 数据源权值。表示该数据源提供的观察值为真实值的概率。权值越大，则表明该数据源提供真实值的可能性越大;权值越小，则表明该数据源提供真实值的可能性越小。各数据源权值总和为 1。

本文中使用的变量定义如表 2 所示。

表 2 文中使用的变量定义

Table 2 Notation of variables

变量	定义
N	Number of Objects
K	Number of Sources
M_i	Number of i -th Object's Attributes
O_k	Collection of Objects from Source s
$\phi_i^{(*)}$	The Observation Set of Object i from All Sources
$\varphi_i^{(k)}$	The Observation Set of Object O_i from Source S_k
τ_i	The Truth Chromosome of i -th Object

1.2 系统架构

本文通过以下五个步骤进行真值发现，如图 1 所示。

a)构建各对象可能值集、观察值染色体、属性染色体、真值染色体和属性类型向量。

b) 根据当前各对象真值染色体和观察值染色体，通过权值计算公式得到各数据源的初始权值。

c)根据步骤 b)得到的各数据源权值，通过改进后的遗传优化算法求得目标函数全局最优解。此解即为当前各数据源权值下的真值集合。

d)迭代重复执行步骤 c)d)，每轮结束后比较前后两次得到的数据源权值，当两者间差异度满足收敛条件时停止算法。

e)根据当前各数据源的权值，确定最终各对象的真值集合。

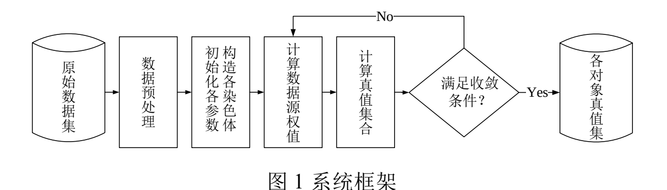


图 1 系统框架

Fig. 1 System framework

2 多属性数据的联合真值发现方法 ATD4MA

2.1 多属性对象各属性间的相关性

对各属性间的相关性进行讨论。对于标称数据，两属性间的相关联系可通过 *Person* χ^2 统计量进行检验。令 (A_i,B_j) 表示属性 A 取值 a_i、属性 B 取值 b_j 的联合事件，χ^2 用下式进行计算：

$$\chi^2=\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \tag{1}$$

其中：o_{ij} 为联合事件 (A_i,B_j) 的观测频度，e_{ij} 为 (A_i,B_j) 的期望频度。e_{ij} 可用下式计算：

$$e_{ij}=\frac{count(A=a_i)\times count(B=b_j)}{n} \tag{2}$$

其中：n 为描述对象的元组个数，$count(A=a_i)$ 为属性 A 上具有值 a_i 的元组个数，$count(B=b_j)$ 为属性 B 上具有值 b_j 的元组个数。

对于数值数据，通过计算两属性的相关系数（*Person* 积矩系数）估计其相关度。可用下式计算：

$$r_{A,B}=\frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n\sigma_A\sigma_B}=\frac{\sum_{i=1}^{n}(a_ib_i)-n\bar{A}\bar{B}}{n\sigma_A\sigma_B} \tag{3}$$

其中：n 为对象的元组个数，a_i 和 b_i 分别是元组 i 在 A 和 B 上的值，$\bar{A}$ 和 $\bar{B}$ 分别是 A 和 B 的均值，σ_A 和 σ_B 分别是 A 和 B 的标准差，$\Sigma(a_ib_i)$ 是 AB 的叉积和。若 $r_{A,B}$ 等于 0，则 A 和 B 是独立的。

抽取电影数据中 1 000 条记录中的演员与电影类型两属性作为数据样本，其中 300 条记录中类型为喜剧。如表 3 所示。

表 3 数据样本的 2*2 相依表

Table 3 2*2 Dependency Table of Data Samples

类型	周星驰	其他演员
喜剧	165	135
其他	32	668

利用式(2)计算每个单元的期望频率，比如单元(周星驰，喜剧)的期望频率为

$$e_{11}=\frac{count(\text{'周星驰'})\times count(\text{'喜剧'})}{n}=\frac{197\times 300}{1000}=59.1$$

同理

$$e_{12}=\frac{197\times 700}{1000}=137.9\,,\ e_{21}=\frac{803\times 300}{1000}=240.9\,,\ e_{22}=\frac{803\times 700}{1000}=562.1$$

利用式(1)可得

$$\chi^2=\frac{(165-59.1)^2}{59.1}+\frac{(32-137.9)^2}{137.9}+\frac{(135-240.9)^2}{240.9}+\frac{(668-562.1)^2}{562.1}=337.59099$$

对于 2*2 的表自由度为 $(2-1)(2-1)=1$。自由度为 1，在 0.001 的置信水平下，拒绝假设的值为 10.828，计算得到的值远大于此值，故可以断定这两个属性间是强相关的。同理数值类型属性通过式(3)也可证明出存在相关性，此处不再赘述。

由上述分析可得，在多属性数据问题的处理上，各属性间的相关性不容忽视。为得到更加精确的结果，在数据源权值计算模型与对象真值求解模型的建立上需要考虑到其相关性并将处理方法加入模型中。而面对与庞大的数据集对各属性两两进行相关性分析会使算法效率大大降低。不同于已提出方法，本文将对象的各个属性视作整体进行建模，确保在计算过程中不破坏对象属性间整体性，从而在避免其相关性影响的前提下大幅度提升了算法的执行效率。

2.2 数据源权值计算

每个数据源为每个对象提供了不同类型的多个属性值，借助遗传算法中染色体的定义，令 $B_{i,m}^{(k)}$ 表示数据源 S_k 为对象 O_i 中第 m 个属性提供的观察值染色体，长度为 L_m。令 $B_i^{(k)}$ 表示数据源 S_k 为对象 O_i 提供的所有观察值染色体，其长度为各属性观察值的长度之和 $\sum_{m=1}^{M}L_m$。则 $B_i^{(k)}$ 中第 l 个元素的取值可表示为

$$B_i^{(k)}[l]=\begin{cases}1, & \phi_i^{(*)}[l]\in\varphi_i^{(k)}\\ 0, & \phi_i^{(*)}[l]\notin\varphi_i^{(k)}\end{cases} \tag{4}$$

其中：$B_i^{(k)}[l]$ 表示数据源 k 提供关于对象 i 的对象观察值染色体中第 l 位基因，$\phi_i^{(*)}[l]$ 表示对象 i 的可能值集中的第 l 个值，$\varphi_i^{(k)}$ 表示数据源 k 提供的关于对象 i 的观察值集。

本文通过数据差异度来衡量两数据间不同部分的大小，即表示两数据间的冲突程度，冲突程度越大，它们之间的差异度越大。染色体中包含多个基因即多个值，在考虑两条染色体间的差异度时不仅要考虑基因取值同为 1 的情况，还应考虑到同为 0 的情况，构造损失函数以表示对象观察值染色体与其真值染色体间差异度。如式(5)所示。

$$\mathbb{C}_{i,m} = \frac{\sum_{l=1}^{L_m} \left|\tau_{i,m}[l] - B_{i,m}^{(k)}[l]\right|}{L_m} \tag{5}$$

其中：$\tau_{i,m}[l]$ 表示对象 i 中第 m 个属性的真值染色体中第 l 个基因的值，L_m 用于消除属性值数量上的不同对结果的影响。由于更高质量的数据源通常会提供更可信的数据，可通过差异度来计算各数据源的权值，此权值的大小描述了该数据源质量的好坏程度。数据源 k 的权值计算公式定义为

$$\omega_k = \frac{1}{\sum_{i=1}^{|o_k|}\sum_{m=1}^{M_i}\mathbb{C}_{i,m}} \tag{6}$$

注意到在权值的计算公式中，ω_k 的计算结果与数据源提供的声明数量 $|O_k|$ 有关。长尾数据的大量存在对数据源质量的评估产生了很大的障碍，导致个别数据源得到非常极端的权值分配，从而影响最后真值计算的准确率。为得到更精确的数据源权值，需将长尾现象加入考虑并消除长尾数据对数据源权值计算产生的影响。

2.3　消除长尾数据对数据源权值计算的影响

在 Web 中并不能保证每个数据源都提供了相同的声明数量，很明显这将导致求和后各数据源被赋予了不公平的权值。本文利用平均数对声明数量的差异进行平滑，即在当前权值计算结果基础上除以该数据源提供声明的数量，从而消除长尾数据源对权值计算带来的不公平性。综上所述，将数据源权值的计算公式改进为

$$\omega_k' = \frac{1}{|o_k|\sum_{i=1}^{|o_k|}\sum_{m=1}^{M_i}\mathbb{C}_{i,m}} \tag{7}$$

对 ω_k' 进行标准归一化处理，同时考虑到方便计算，对原公式取对数，数据源权值的计算公式最终形式为

$$\omega_k'' = -\ln\left(\frac{\omega_k'}{\sum_{p=1}^{K}\omega_p'}\right) \tag{8}$$

每次迭代后为评价数据源权值的变化，同样可由差异度来衡量前后变化大小。两数据源权值间的差异度计算公式如式(9)所示，其中 w_i 表示本次迭代后数据源 i 的权值，w_i' 表示上次迭代后数据源 i 的权值。

$$d = \sum_{i=1}^{K}\sqrt{(w_i - w_i')^2} \tag{9}$$

2.4　对传统遗传优化算法的改进

在 2.1 节已经证明了对象各属性间可能存在相关性，本文将一条包含多个属性的记录视为一个整体进行处理，如此一来在真值计算的迭代过程中可以确保该记录的完整性，避免再加入属性间相关性的修正因子。同时本文方法力求实现单真值发现问题与多真值发现问题的联合求解，即一条记录中又可能同时包含单值属性与多值属性。需要一种可变策略，分别适用于单值属性和多值属性的操作，即一种即能满足各属性值随意变动但此变动必须处于一定范围内的数据载体。

对象真值计算的目标为各对象真值结果集与各数据源提供该对象观察值之间差异度加权和达到最小，故目标函数可设定为

$$\min\left(\sum_{k=1}^{K}\omega_k''\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbb{C}_{i,m}\right) \tag{10}$$

由于在 2.3 节中加入了 $|O_k|$ 平滑对象声明数量的差异，此因子必定导致该目标函数为一个非凸非凹函数，图像会中出现多拐点的现象，即存在多个局部最优解，不能使用传统对于凸或凹的目标函数的求解方法。

根据上述两点分析，首先借助遗传算法对群体初始化算法和染色体的基本动作分别进行改进，即可满足上述分析中的要求。每次迭代染色体作为一个整体带入进行计算，保证了一条记录中各属性间的完整性，避免破坏属性间相关性而带来影响。对染色体动作交叉变异进行针对问题的特点进行修改，可保证单值属性和多值属性中值的跳动在一定范围内。之后构造属性类型向量保证结果中属性值数量的正确性，防止单值属性出现多值的错误情况。最后通过遗传算法解决目标函数存在多个局部最优解的问题，按照优胜劣汰的原则对染色体种群进行逐代演化从而逐步逼近全局最优解。详细如下：

a)群体初始化算法的改进。将每条数据记录以染色体的形式表示。随机生成 G 条染色体，其中每条染色体为一个解，其长度 L_n 为该对象可能值集的长度。在构造群体中染色体时，为保证解的正确性规定单值类型的属性观察值染色体中只允许出现某个基因为 1，而不能是多个，故定义属性类型向量 $Flag$ 用来表示各属性的类型，其长度为对象中属性的个数。其中单值类型属性在对应的位置标记为 0，多值类型标记为 1。同时定义长度向量 Len 用来存放对象可能值集染色体中各属性对应子染色体的长度，以保证之后赋予各属性正确的值数量。具体流程如算法 1 所示。

算法 1　改进后的群体初始化算法 $InitP(Flag, M, L_n, Len)$

输入：类型标记向量 $Flag$、群体大小 M、染色体长度 L_n、各属性子染色体长度向量 Len。

输出：真值向量群体 P。

for m = 0 to M
　定义长度为 L_n 的向量 $\overline{T_m}$；
　for n = 0 to $Flag.count$
　　定义长度为 $Len[n]$ 的全 0 向量 $\overline{t_n}$；
　　if $Flag[i]$ == 0 then
　　　随机生成 0 到 $Len[n]$ 内的一个整数 r；
　　　标记 $\overline{t_n}$ 中第 r 位为 1;
　　else　$\overline{t_n}$ 中的每一位分别随机标记为 0 或 1;
　　end if
　end for
　将生成的 $\overline{t_0}$ … $\overline{t_n}$ 合并赋予 $\overline{T_m}$；
　将 $\overline{T_m}$ 插入群体 P；
end for
return 群体 P { $\overline{T_0}, \overline{T_1}$,?..$\overline{T_M}$ };

b)染色体评价。遍历当前群体中的各染色体，分别代入目标函数计算适应度，即计算结果的好坏程度。

c)选择运算的改进 $Select(O)$。将适应度最好的前 $n/10$ 条染色体选择出来，其中 n 为对象的记录总个数，直接作为下一代群体中的染色体。为防止算法出现早熟现象，在 $[n - n/10, n]$ 范围内随机挑选 3 条记录直接进入下一代群体中。

d)交叉运算的改进 $Swap(O)$。交叉运算使得算法搜索全局最优解能力得以飞跃性的提高，传统的交叉算法将两个染色体的部分结构加以替换重组而生成新的染色体，期望将有益基因组合在一起。

染色体中属性子染色体达到问题定义的最小粒度级，针对本问题的特殊性本文在传统交叉操作上做了改进：在染色体交叉时，将单值属性对应的部分按该属性对应的块进行整体互换后形成新的染色体。假设某单值属性可能值集长度为 4，交叉过程如图 2 所示，其中虚线框部分为某属性的观察值染色体。

多值属性在交叉操作时，采用多点交叉，交叉位数定为此属性对应的长度 L_m 除以 2 后向下取整，即随机选取 n 位进行多点交叉，其中 $n = \frac{L_m}{2}$。如图 3 所示。

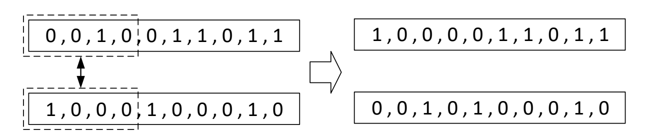


图 2 单值属性交叉操作

Fig. 2　Cross operation of single-valued attribute

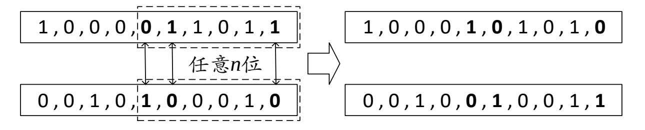


图 3 多值属性交叉操作

Fig. 3　Cross operation of multi-valued attribute

e)变异运算的改进 $Mutate(O)$ 。

对群体中各染色体内某些基因座上的基因值进行变动。变异运算使得算法具有局部的随机搜索能力，当遗传算法通过交叉运算后已接近最优解邻域时，利用变异运算的这种局部随机搜索能力可以加速向最优解收敛。

针对本问题本文将单值属性的变异操作改为跳位操作，即染色体中单值属性对应部分中的‘1’随机跳至该部分中的另一位置。假设某单属性可能值集长度为 4，变异过程如图 4 所示，其中虚线框部分为某属性观察值染色体。

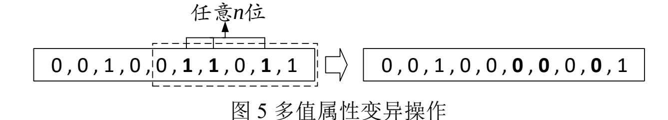


图 4 单值属性变异操作

Fig. 4　Mutate operation of single-valued attribute

多值属性在变异操作时随机选取该属性对应范围内 n 位进行突变，即逢 0 变 1，逢 1 变 0，其中 $n=\frac{L_m}{2}$ 。如图 5 所示。

任意n位

0,0,1,0,0,1,1,0,1,1　　0,0,1,0,0,0,0,0,0,1

图 5 多值属性变异操作

Fig. 5　Mutate operation of multi-valued attribute

反复执行步骤 b）~e)，直到某一代群体中的目标函数最优值与上一群体中的目标函数最优值连续出现三次相等情况时停止算法，此时导致最优目标函数值的染色体，即为当前权值下该对象的最优真值染色体，实现了单真值与多真值的联合计算。改进后群体初始化算法省去了遗传算法中低效的编码与解码操作；选择操作避免了算法容易早熟的缺陷；交叉与变异操作省去了交叉率与变异率的设定，以属性为最小粒度使算法更快至收敛状态；回避传统遗传算法缺陷的同时提升了计算效率。

2.5　多属性数据的联合真值发现方法（ATD4MA）

文献[13]已经给出了联合迭代机制推导数据源质量与真值集合方法有效性的严格证明，本文在真值计算上也采取这种机制。首先初始化一个真值向量进行各数据源的可信度权值计算，之后利用计算得到的数据源权值进行各对象真值集合的计算。每次迭代都会更新各数据源的权重与各对象的真值集合，直到本次计算与上次计算得到的各数据源权值间的差异度在一定范围内，则认为算法达到收敛状态，可由式(9)来评价此差异度。此时根据当前各数据源权重，计算得到各对象的真值集合。具体流程如算法 2 所示。

算法 2　基于多属性数据的联合真值发现方法（ATD4MA）

输入：所有数据源提供的冲突数据集。

输出：各数据源质量 W 与各对象的真值集合 T 。

```
初始化各对象的真值染色体集 V={τ1,τ2,…τN}；
计算各数据源权值，更新 W={w1,w2,…wK}；
do
  W'=W；
  for i=1 to N
     根据对象中属性类型更新类型向量 Flag；
     更新子染色体长度向量 Len；
     定义群体大小 M 与染色体长度 Ln；
     P=InitP(Flag,M,Ln,Len)；
     do
        P=Mutate(Swap(Select(P)))；
        将导致最优适应度的染色体赋予 temp；
        if vi=temp then break;
        else vi=temp；
        end if ；
     while(1)
     τi=vi；
  end for
  将 V 代入式(8)计算各数据源的权重，更新 W；
  if W 与 W' 间差异度满足收敛条件 then break;
  end if ；
while(1)
根据 V 与 Φ* 得出最终的真值结果集 T；
return T,W；
```

假设算法运行至收敛时迭代的次数为 K ，则算法的时间复杂度为 $O(NMK)$ 。

3　实验与分析

3.1　数据集与实验环境

在两个真实的数据集上进行实验。

a)电影数据集。从 Web 中爬取近 10 年豆瓣评分高于 6.0 的电影数据。经过预处理后的数据集包含 405 部电影，来自于 92 个网站，共 3 719 条记录。选取其中 50 部电影，对海报上信息进行人工确认后作为基准数据集。

b)书籍数据集。从 Web 中爬取书籍数据集，包含了来自 450 个数据源的 2 245 本图书，22 972 条冲突数据记录。在其中随机选择 50 本图书，通过其封面信息对其作者信息进行人工确认后作为基准数据集。

实验运行环境为：Intel(R) CoreTM i5-7300HQ CPU@ 2.50 GHz (4 CPUs)处理器、16 GB 内存、Windows 10 操作系统，数据库为 SQL Server 2012，所有算法均使用 MATLAB 语言实现。

3.2　评价指标

真值发现的目标即在冲突数据集中找到最准确、最完整的真值集合。为评价结果的准确性和完整性，本文通过三个指标来评价本文所提方法。

a)查准率(*precision*)。衡量计算得到的对象真值相比于该对象实际真值的准确率。假设某对象的真实数据集中包含 n 个值，计算得到的真值结果中包含 m 个值，其中有 p 个值属于真实数据集，表示为 $pre=\frac{p}{max(n,m)}$ 。

b)查全率(*recall*)。衡量计算得到的对象真值在该对象的实际真值集中所占比率大小。假设某对象的真实数据集中包含 n 个值，计算结果中包含 m 个真实的值。表示为 $rec=\frac{m}{n}$ 。

c)调和平均数 $F-Score$ (*harmonic mean*)。衡量结果整体水

平，即查准率和查全率的调和平均值。表示为 $F_s = \frac{2\times pre\times rec}{pre+rec}$ 。

3.3 算法收敛的判定

对 ATD4MA 算法收敛的条件进行讨论。当本次计算与上次计算得到的各数据源权值间差异度在一定范围内时，认为算法达到收敛状态，停止迭代。分别对电影和书籍两数据集进行 10 次迭代，计算每次迭代后的差异度。实验结果如图 6 所示，其中横坐标表示迭代次数，纵坐标表示差异度。

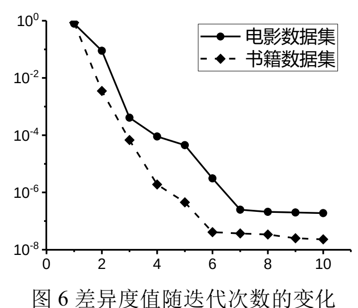


图 6 差异度值随迭代次数的变化

Fig. 6 The dis-value of each iteration

从图中可看出，电影数据集在经过 7 次迭代后，可达到收敛状态，此时差异度约为 10^{-6} 。书籍数据集在经过 5 次迭代后，可达到收敛状态，差异度约为 10^{-7} 。故在两数据集上，当算法执行至差异度分别小于 10^{-6} 与 10^{-7} 时，可判定算法满足收敛条件并停止迭代。同时，本实验结果也证明了本文提出算法可达到快速收敛的效果。

3.4 ATD4MA 与传统算法的对比

实验中选取了如下三种方法与本文算法进行比较：

a)Voting。采取投票机制，以各对象为单位，各对象选择各数据源中出现最多次数的描述值为该对象的真值。

b)TruthFinder。文献[1]中提出的单真值发现方法，考虑到了数据源可信性的分配问题。

c)Mtruth。文献[12]中提出的多真值发现方法，又可分为枚举和贪心两种策略进行。枚举策略较贪心策略会耗费更多的时间，但枚举策略的准确性高于贪心策略。本次实验中选择与其中的枚举策略进行比较。

为使各方法之间具有可比性，实验首先在对象只包含一个属性的情况下分别与单真值发现方法和多真值发现方法进行比较，以证明 ATD4MA 同样可胜任之前提出方法所解决的问题。之后在对象包含多个属性的情况下与个算法进行比较，当面对多个单值属性与多个多值属性时，Voting、TruthFinder 和 MTruth 采取属性拆分后分别进行计算。当面对单值属性与多值属性同时存在情况时，将单真值发现方法 TruthFinder 与 MTruth 结合，将多个属性拆分后分别进行计算。而 ATD4MA 面对上述情况可直接进行整体计算，如此进行比较以突出本文论点。

3.4.1 数据包含单个属性时各算法的比较

当对象只包含单个属性时，ATD4MA 相比于单真值发现方法 Voting 和 TruthFinder，实验中选取电影数据中电影时长和书籍数据中售价两个单真值属性，表现分别如图 7(a)(b)所示。相比于多真值发现方法 MTruth，实验中选取电影演员和书籍作者两个多真值属性。其中 Voting 和 TruthFinder 需设定一个阈值，即属性值为真的概率大于该阈值时判定该属性值为真，实验中将该阈值设定为 0.75。两数据集下各算法的表现分别如图 7(c)(d)所示。

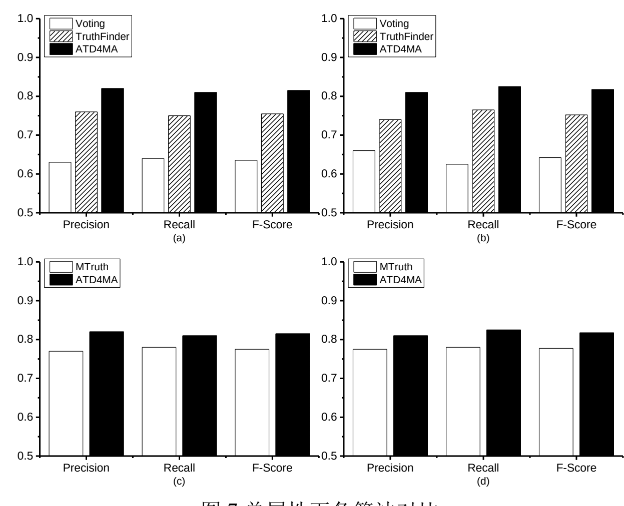


图 7 单属性下各算法对比

Fig. 7 Comparison of each algorithms under single attribute

由于 ATD4MA 考虑到了长尾数据对计算结果的影响，相比于单真值发现方法 Voting 和 TruthFinder 表现出较差的效果，并在多值属性数据集的表现高于 MTruth。由此说明本文方法分别适用于单真值与多真值数据集的真值计算，相比于传统算法表现出更佳的效果。

3.4.2 数据包含多个属性时各算法的比较

当对象同时包含单值属性与多值属性时，传统方法采取属性拆分后分别处理的方式。分别与传统单真值方法与多真值方法进行比较。相比于单真值发现方法 Voting 和 TruthFinder，实验选取电影数据中导演与时长两个单值属性，表现如图 8(a)所示。相比于多真值发现方法 MTruth，实验选取书籍数据中类别和作者两个多值属性，表现如图 8(b)所示。之后选取电影数据中时长和演员两属性，书籍数据中售价和作者两属性，即一个单值属性和一个多值属性，用单值发现方法 TruthFinder 与多值发现方法 MTruth 分别处理，电影数据集和书籍数据集中各方法表现分别如图 8(c)(d)所示。

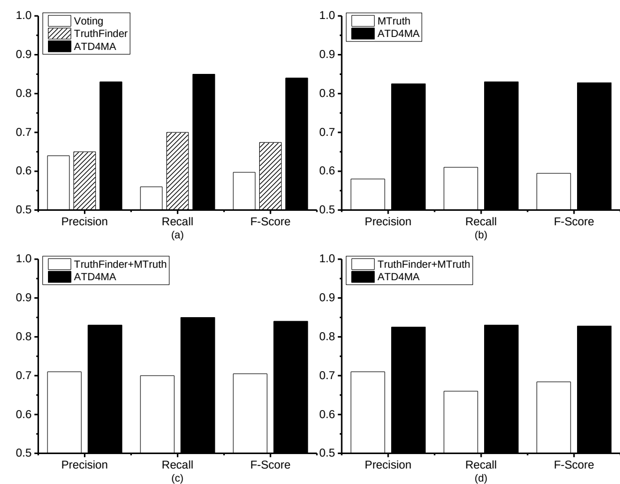


图 8 多属性下各算法对比

Fig. 8 Comparison of algorithms under multiple attribute

由图可得，在对象包含多个单值于多值属性的情况下，传统方法只能处理单个属性，因只能采取拆分处理，破坏了各属性间相关性从而导致不精确的数据源权值和真值集，故传统方法表现出较差效果。当对象同时包含单值与多值属性时，将属性进行拆分后分别通过单真值发现与多真值发现方法计算也因破坏了属性相关性同样表现出较差的效果。而 ATD4MA 对属性进行整体联合处理，没有破坏对象各属性间的相关性，表现出了较好的效果。

综上，两个数据集上的实验证明了本文提出的方法在对象同时包含多个单值或多值属性的情况下，可有效地计算出精确的真值集合，而以往提出的方法均无法应对此种情况下的真值发现问题。本文方法在之前方法基础上考虑到数据中普遍存在的严重影响数据源权值精确度的长尾现象，借助遗传算法优点的同时避开其缺点对问题进行建模，相比于之前方法得到更为精确的计算结果。因遗传算法的自身特点，该方法在算法执行时间上相比于之前方法会有所增加。考虑到真值发现是一个一次性的过程，牺牲一定的时间换取准确性、查全性方面的大幅度提升是值得的。

4 结束语

之前提出的真值发现方法，只能处理对象包含一个属性时的真值发现，例如书籍的作者，电影的导演。在当今大数据时代下完全不能满足数据的处理要求。当需要处理对象的多个属性时，之前方法显得力不从心。本文提出可用于多个属性同时进行真值发现的方法 ATD4MA，借助染色体作为数据载体，对遗传优化算法的群体初始化算法和染色体的基本动作分别进行改进改进，取其优避其劣，克服了模型求解陷入局部最优解的问题，在不破坏属性间相关性的情况下实现了单值与多值属性的联合真值发现计算，为真值发现研究领域的一大突破。ATD4MA 同适用于之前提出的真值发现方法所解决的问题并表现出更佳的效果。最后在两组真实数据集上的实验证明了本文提出算法的有效性。接下来的进一步工作，将致力于研究可适用于动态数据流的高效真值发现方法与算法执行效率方面的改进与提升。

参考文献：

[1] Yin Xiaoxin, Han Jiawei, Yu Shilun. Truth discovery with multiple conflicting information providers on the Web [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 796-808.

[2] Galland A, Abiteboul S, Marian A, *et al.* Corroborating information from disagreeing views [C]//Proc of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 131-140.

[3] Dong Xin, Berti-Equille L, Srivastava D. Integrating conflicting data: the role of source dependence [J]. Proc of the VLDB Endowment, 2018, 2(1): 550-561.

[4] Dong Xin, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world [J]. Proc of the VLDB Endowment, 2009, 2(1): 562-573.

[5] 余东，申德荣，寇月，等. 面向 Web 数据集成的真值发现算法 [J]. 小型微计算机系统, 2016, 37(8): 1633-1638. (Yu Dong, Shen Derong, Kou Yue, *et al.* Web data integration oriented truth discovery algorithms [J]. Journal of Chinese Computer Systems, 2016, 37(8): 1633-1638.)

[6] Li Xian, Dong Xin, Lyons K B, *et al.* Scaling up copy detection [C]// Proc of the 31st International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2015: 89-100.

[7] Li Qi, Li Yaliang, Gao Jing, *et al.* A confidence-aware approach for truth discovery on long-tail data [J]. Proc of the VLDB Endowment, 2014, 8(4): 425-436.

[8] Lin Xueling, Chen Lei. Domain-Aware Multi-Truth discovery from conflicting Sources [J]. Proc of the VLDB Endowment, 2018, 11(5): 635-647 2150-8097.

[9] 马如霞，孟小峰. 基于数据源分类可信性的真值发现方法研究 [J]. 计算机研究与发展，2015, 52(9): 1931-1940. (Ma Ruxia, Meng Xiaofeng. Credibility of the discovery of the true value based on the data source classification [J]. Journal of Computer Research and Development, 2015, 52(9): 1931-1940.)

[10] Pochampally R, Sarma A D, Dong Xin, *et al.* Fusing data with correlations [C]// Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014: 433-444.

[11] Zhao Bo, Rubinstein B I P, Gemmell J, *et al.* A Bayesian approach to discovery truth from conflicting sources for data integration [J]. Proc of the VLDB Endowment, 2012, 5(6): 550-561.

[12] 马如霞，孟小峰，王璐，等. MTtuths: Web 信息多真值发现方法研究 [J]. 计算机研究与发展, 2016, 53(12): 2858-2866. (Ma Ruxia, Meng Xiaofeng, Wang Lu, *et al.* MTruth: An approach of multiple truths finding from web information [J]. Journal of Computer Research and Development, 2016, 53(12): 2858-2866.)

[13] Xiao Houping, Gao Jing, Wang Zhaoran, *et al.* A truth discovery approach with theoretical guarantee [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1925-1934.

[14] Li Qi, Li Yaliang, Gao Jing, *et al.* Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014: 1187-1198.